A Fine Rule From a Brutish World?

An Experiment on Endogenous Punishment Institution and Trust*

Huojun Sun†,
University of Bologna

Maria Bigoni‡
University of Bologna & IZA

Abstract. By means of a laboratory experiment, we study the impact of the endogenous adoption of a collective punishment mechanism within a one-shot binary trust game. The experiment comprises three games. In the first one, the only equilibrium strategy is not to trust, and not to reciprocate. In the second we exogenously introduce a sanctioning rule that imposes on untrustworthy second-movers a penalty proportional to the number of those who reciprocate trust. This generates a second equilibrium where everybody trusts and reciprocates. In the third game, the collective punishment mechanism is adopted through majority-voting. In line with the theory, we find that the exogenous introduction of the punishment mechanism significantly increases trustworthiness, and to a lesser extent also trust. However, in the third game the majority of subjects vote against it: subjects seem to be unable to endogenously adopt an institution which, when exogenously imposed, proves to be efficiency enhancing.

Keywords: Coordination, Majority Voting, Social Sanctions, Trust Game

JEL codes: C72, C92, D72

* We thank Stefania Bortolotti, Marco Casari, Davide Dragone, Diego Gambetta and Paolo Vanin for insightful comments. This paper also benefited from comments received by seminar participants at the University of Bologna, the European University Institute, and the European School on New Institutional Economics. The usual disclaimer applies. We gratefully acknowledge financial support from the Law and Economics Research Center of Zhejiang University (RG201310004), and from the Italian Ministry of Education (grant FIRB-Futuro in Ricerca no. RBFR084L83).
† Died January 11, 2018.
‡ Corresponding Author. Department of Economics, University of Bologna, piazza Scaravilli 2, Bologna, 40126, Italy; Telephone: +39 051 209 8134. E-mail: maria.bigoni@unibo.it

1. Introduction

A well-functioning and impartial legal system largely enhances societal trust, thereby promoting trade and economic development (Algan and Cahuc, 2013; Guiso, et al., 2008; Tabellini, 2008). Better enforcement can increase the likelihood of contract performance, naturally stimulating all manner of reliance investments that have specific value in the contractual relationship (Polinsky and Shavell, 2008).[1] Nearly half of the world's governments, however, fail to provide a sufficiently strong system of contract enforcement (Leeson and Williamson, 2009), and even abuse their authority to engage in profit-seeking punishment, which is detrimental to the country's economic performance (Xiao, 2013). Therefore, it becomes of paramount importance to understand how people who lack the protection of an effective legal environment can establish private-order institutions (or norms) to facilitate mutually advantageous exchanges.

Here, we study this issue by means of a laboratory experiment based on a model proposed by Anderlini and Terlizzese (2017), who theoretically study the introduction of a social norm into a standard contractual relationship, by letting the promisor's behavior be constrained by the average behavior of other promisors in a society. More specifically, the model represents a bilateral contractual relationship in the absence of contract enforcement as a one-shot binary trust game. Think for instance of an investor and an agent, strangers to each other. The investor lends some money to the agent, who makes an investment, and this investment generates a surplus proportional to the invested sum. The agent then decides whether to cheat and keep the entire surplus, or to share it with the investor. In the model, cheating entails a cost characterized by two components: a "moral" component – which is idiosyncratic and depends on the exogenously given "type" of the agent[2] – and a "norm-driven" component – which is socially determined and common to all agents, and depends on the total number

[1] Introducing a third-party intervention into an investment game, Charness et al. (2008) experimentally reveal that the incentives (i.e. sanctions or rewards) implemented by an independent third-party significantly increase trust and trustworthiness in the investment game.

[2] Previous experimental studies have revealed that individuals involved in social dilemmas are heterogeneous in terms of social preferences (Blanco et al., 2011). Anderlini and Terlizzese (2017) assume that there are two types of agents, high-type and low-type agents, who differ in their preference for honesty and the magnitude of the psychological cost they suffer when abusing their partner's trust.

of transactions in a society that go through without cheating. In other words, the stronger the norm of trustworthiness in a society, the higher the cost of cheating for the agents. Anderlini and Terlizzese (2017) note that the norm-driven component of the cheating cost can be interpreted as reflecting psychological remorse when the agent's action deviates from average behavior (Huang and Wu, 1994), or as resulting from a collective punishment mechanism, a form of stigma, whose effectiveness depends on the average behavior. Our experimental design adopts the second perspective, potentially inflicting a social sanction on the dishonest agents. The introduction of this norm-driven component of the cost of cheating transforms the trust game into a coordination game with high-trust and low-trust equilibria, which are Pareto-ranked.

Existing experimental evidence indicates that norms of trustworthiness may differ across societies (Buchan et al., 2002), and such a difference might affect individual behavior, inducing the emergence of one or other of the equilibria. The issue of how social norms emerge, however, remains largely unexplored. Anderlini and Terlizzese (2017) assume that the "social sensitivity" to the norm-driven component of the cheating cost is exogenously given. In this study we take a further step and investigate the effects of the endogenous adoption of a collective punishment mechanism whose intensity is proportional to the strength of the norm of trustworthiness in society. In one treatment, the adoption of the punishment mechanism is based on majority voting.[3] This allows us to study whether the endogenous adoption of the norm can help a society to coordinate on an efficient equilibrium, characterized by high levels of trust and trustworthiness. Starting from a simplified version of Anderlini and Terlizzese's model, we theoretically show that most subjects, regardless of their moral cost of cheating and expectations, should vote in favor of the punishment mechanism, hence this mechanism will be endogenously introduced. Consequently, a majority vote in favor of collective punishment cannot be interpreted as a signal of subjects' intentions, and it should not matter whether collective punishment is exogenously imposed or endogenously adopted. This theoretical prediction contrasts with the findings of recent experimental studies, which revealed that the endogenous adoption of institutions induces higher cooperation levels in social dilemma situations, relative to the case in which

[3] In real world, we rarely observe that the norm is established through a voting mechanism. However, people in a community could publicly express their attitudes towards a specific norm (Kadens and Young, 2013). Therefore, we use the voting mechanism as a simple way to capture the essential dimension of the public expression of the norm.

the same institutions are exogenously implemented; scholars refer to this phenomenon as "the dividend of democracy" (Andreoni and Gee, 2012; Dal Bo et al., 2010; Markussen et al., 2014; Sutter et al., 2010; Tyran and Feld, 2006).[4]

The theoretical model informs our empirical analysis, which is based on a laboratory experiment. In our experiment, each subject plays three one-shot games with three different partners. The first game is a standard binary trust game. In the second game, a collective punishment mechanism is exogenously introduced, under which cheating is sanctioned with a severity that depends on the trustworthiness of the others. In the third part of the experiment subjects must choose whether to play according to the rules of the first, or of the second game, by means of a majority voting mechanism. To reduce the risk of spillover effects, the outcomes of these three games are not revealed to the subjects until the end of the session. In half of the sessions the sequence of the first and the second game is reversed, to control for possible order effects. This design allows us to test whether subjects are willing to opt for having a collective punishment mechanism in place, and to study how the endogenous adoption of such a mechanism affects individual beliefs and behavior.

We report four main findings. First, in line with the model, we find that the introduction of collective punishment induces a significant increase in the levels of trustworthiness, and to a lesser extent also of trust. Second, the endogenous introduction of the punishment mechanism by means of a majority-voting rule does not significantly change behavior, with respect to what is observed when the mechanism is exogenously imposed. Third, in contrast with our theoretical predictions, not all subjects seem to be able to anticipate the change in behavior induced by the introduction of collective punishment, and most of them vote against it. We also find that subjects with higher cognitive abilities and with a background in statistics are more likely to vote in favor of the punishment mechanism. To study whether the decision not to vote for the punishment mechanism depends on subjects' inability to anticipate its consequences, in an additional treatment we provide information about the aggregate

[4] Vollan et al. (2013) replicate Tyran and Feld's (2006) study using a sample of Chinese people. They observe that the cooperation rate is higher under an exogenously imposed institution than under a democratically selected rule. Their analyses show that this result is mainly driven by the fact that the Chinese culture attributes a high importance to obeying authorities.

behavior with and without collective punishment; we find that this does not increase the number of subjects who vote in favor of the punishment mechanism.

The paper proceeds as follows: Section 2 discusses how our work relates to the existing literature. Section 3 presents our theoretical model and testable predictions; Section 4 describes the experimental design and procedures; Section 5 illustrates the main results of the experiments; Section 6 concludes.

2. Related Literature

Our paper builds upon a considerable number of studies on the effects of informal institutional arrangements on individual behavior in social dilemma situations, in the absence of a powerful state (Ostrom, 1990). A variety of decentralized governance institutions have emerged in remarkably diverse environments (Bernstein, 1992, 2001; Greif, 2006).

In early trade, Greif (1989, 1993) portrays a well-defined and cohesive group based on Jewish religion and family origins in the Maghreb, the "Maghribi traders" who engage in long-distance, large-scale trading across the whole Muslim Mediterranean. Lacking effective legal institutions, these merchants rely on informal sanctions based on collective relationships within an exclusive coalition. Members of the Maghribi traders' coalition always recruit agents from their own coalition, convey information about their agent's misbehavior swiftly to other members, and collectively ostracize agents who abused their principal's trust, thereby successfully resolving the problem of commitment in one-shot bilateral contractual relationships, even in the absence of binding contracts. Similar social sanction institutions also proved to work well in Mexican California before the time of the gold rush in 1848-1949 (Clay, 1997; Clay and Wright, 2005) and in the practice of group lending in the developing countries (Besley and Coate, 1995).

These anthropological studies on informal sanctioning institutions emphasize the role of information-sharing among the investors in regulating the agents' behavior.[5] By contrast, our research

[5] In Kimbrough and Rubin (2015), subjects play the trust game under a highly anonymous set-up, where the investors only know the group identity of their agents. When the investors can share their transaction experience with other investors, the groups with high percentages of dishonest agents are collectively boycotted, which secures the high efficiency of the market.

adopts an alternative approach: in our set-up, to gain the investors' trust, agents are allowed to adopt a collective punishment mechanism whose severity depends on the average behavior of all agents' in the society. Therefore, the effectiveness of our mechanism relies on the agents' and the investors' beliefs, rather than on information-sharing.

Secondly, our paper is also related to the literature on expressive law (Cooter, 1998; McAdams, 2000a, 2000b; Posner, 1998, 2000). The classic "law and economics" approach focuses on deterrence: a law enforced by a sanction increases the expected costs of the illegal activity and thereby induces compliance (Becker, 1968; Polinsky and Shavell, 2000). This view can hardly explain why most people obey legal rules even in a situation where they could improve their material payoffs if they violate an obligation (Tyler, 1990). According to the expressive law theories, there are several potential explanations: legal rules may affect individual preferences by making a normative prohibition more salient, act as coordination devices, or work as a form of "cheap talk".[6]

These theories have increasingly gained momentum among theoretical scholars. However, only a handful of experimental studies have examined how mild rules influence individual behavior (Bohnet and Cooter, 2003; Galbiati and Vertova, 2008; Masclet et al., 2003; McAdams and Nadler, 2005; Noussair and Tucker, 2005; Samek and Sheremeta, 2014; Tyran and Feld, 2006). Our experimental study contributes to this literature in two aspects. First, the social sanction in our experiment is not always a deterrent but works only if the majority behaves honestly. Therefore, socially shared beliefs are crucial to affect individual behavior. Second, instead of being announced by a powerful authority, the rule in our paper is determined by a voting mechanism, which enhances its legitimacy and may influence individual behavior through changing people's preferences or coordinating their beliefs. Our study is also related to the experimental literature on the trust game with punishment (Fehr and Rockenbach, 2003; de Quervain et al., 2004; Vollan, 2011), however, it departs substantially from it, in

[6] Despite being "cheap", some forms of talk, especially announced by a powerful authority or determined by a majority voting mechanism, have been found to actually coordinate individuals' behavior in social dilemma situations. For example, in Kamei (2014) subjects are more likely to contribute to cooperation in the public good game when a mild sanction rule is collectively selected even without altering the equilibrium of full free riding.

that the activation and size of the punishment in our case depends on the behavior of the whole society, and not on the individual decision of a trustor who may sanction an untrustworthy trustee.

A closer relation emerges between our work and the literature concerning the endogenous adoption of institutions. Recent experimental studies have revealed that an institution established endogenously (e.g. through a voting mechanism) can induce higher cooperation levels in social dilemmas, compared to the same institution implemented exogenously on an otherwise identical group (Dal Bo et al., 2010; Markussen et al., 2014; Sutter et al., 2010; Tyran and Feld, 2006).

Broadly speaking, there are two approaches to endogenous institution formation. Under the first approach, groups are fixed, and members of each group are asked to vote for a specific scheme or to choose one from a broad menu of schemes (Dal Bo et al, 2017; Kosfeld et al., 2009; Sutter et al., 2010). Previous experimental results indicate that the endogenous adoption of informal sanctioning (Tyran and Feld, 2006; Ertan et al., 2009) or rewarding (Sutter et al., 2010) institutions largely enhances the levels of cooperation, relative to the case in which the same institutions are imposed exogenously. In addition, subjects tend to converge on the most efficient institutions as they gain experience over a course of multiple votes (Putterman et al., 2011). Our results are at odds with this, but they are in line with the recent findings by Dal Bo et al. (2017), who illustrate how people might end up voting for the least efficient institution, because they fail to anticipate the impact institutions have on others' behavior.

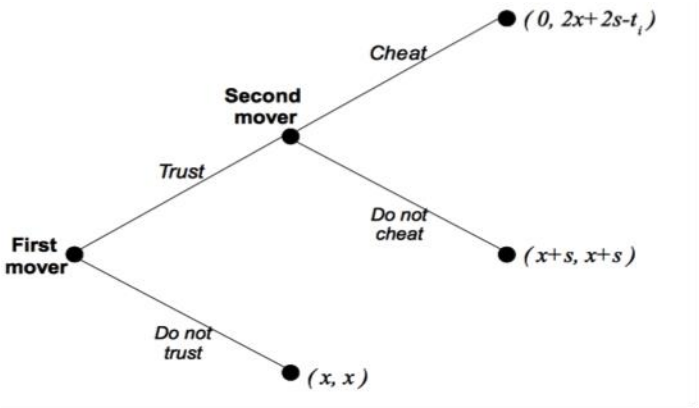
The second approach is the "voting by feet" mechanism in open communities (Gurerk et al., 2006, 2014; Fehr and Williams, 2013) where subjects can choose between different institutions and endogenously form groups with other members who also select the same institution. They find that prosocial individuals adopting efficient punishment institutions under endogenous selection quickly establish a cooperative culture. These institutions increasingly attract other types of subjects to migrate to these more cooperative groups and to comply with the prevailing norms. Therefore, endogenously chosen institutions induce the whole group to coordinate on high cooperation levels, so that in practice there is little or no need to recur to punishment.

Most experimental papers on endogenous formation of institutions are based on the framework of public good games. To the best of our knowledge, no existing empirical research addresses the effect of endogenous adoption of social sanction mechanisms on individual behavior in the trust game. Compared to the previous studies, our within-subject design without feedback across games allows us to identify the important role of ex-ante beliefs of subjects in equilibrium selection. Furthermore, since subjects are exposed to the trust game with and without the collective punishment mechanism before voting for the preferred rule governing their interactions, we can investigate how different experiences of the effects of collective punishment affect individual's voting behavior. Finally, in line with what is argued by Markussen et al. (2014), that "the dividend of democracy" is driven by the signaling function of voting which promotes coordination on high-contribution outcomes, our design also allows us to test whether the endogenous adoption of the punishment mechanism could be taken as signal of the general willingness to coordinate on a high trust and high trustworthiness equilibrium.

3. Theoretical framework

In this section, we present the theoretical model that informs our experimental design, and derive the predictions, which will be empirically tested in Section 5. As a baseline situation, we consider the binary investment (or trust) game depicted in Figure 1. Each player is initially given an endowment $x > 0$. The first mover decides whether to trust the second mover or not. If she chooses not to trust her partner, both of them keep their endowments and leave the transaction. If instead she chooses to trust and transfers her endowment, the second mover efficiently invests the money he received, together with his own endowment, to generate a total of $2x + 2s$, with $s > 0$. The second mover now has to choose whether to cheat on the first mover, and keep the entire amount leaving the first mover with nothing, or to split it equally with her, so that each party gets $x + s$. We further assume that, in the society, all players face equal chances of playing the game in the role of the first or second mover.

Figure 1: the basic trust game.



Following the Anderlini and Terlizzese's (2017) approach, we assume that there are two types of players in the society, "high" (*H*) and "low" (*L*). *H*-type players have a preference for honesty and suffer a psychological cost $t_H > 0$ when abusing their partner's trust, and the idiosyncratic cost of cheating for the *H*-type players is so high that they will never cheat: $t_H > x + s$. *L*-type players instead are only interested in (expected) monetary payoffs (i.e. $t_L = 0$), so they will always cheat when in the role of second movers. For simplicity, we also assume that players are risk neutral.

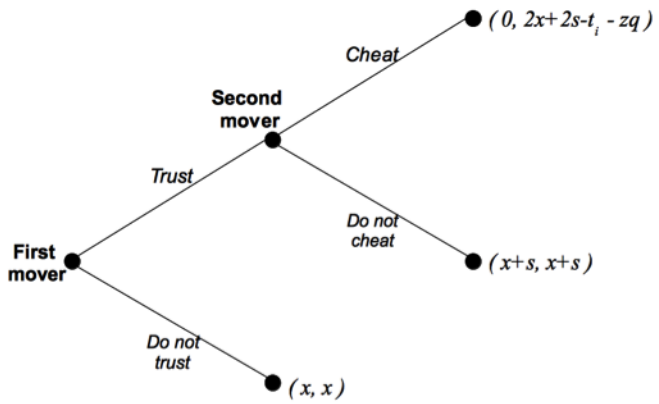
At the beginning of the stage game, all players are randomly assigned to the role of first or second mover and matched in pairs. Players choose their strategy before knowing their role, and the strategy determines their action both as the first and as the second mover. Let p represent the proportion of *H*-type players in the society, which is assumed to be common knowledge. Regardless of his type, as a first mover a player will choose to trust if the expected payoff from trusting $p(x + s)$ is larger than the outside option x, that is if $p > \frac{x}{x+s}$. Let us denote this threshold Θ.

3.1. A collective punishment mechanism

Now consider the introduction of a collective punishment mechanism into the trust game, as depicted in **Figure 2**. In this new game, besides possibly suffering the psychological cost t_i, the player who cheats faces the risk of social stigma. This potential punishment zq depends on two elements: the fraction q of other players who do not cheat, and the strength of the sanction they will impose on cheaters, which is denoted by z and is exogenously given. The behavior of the *H*-type players as second movers is not affected by the sanction, as they would never cheat, in any case. The behavior of the *L*-type players

instead might change, as they may choose not to cheat either, if $q_i \geq \frac{x+s}{z}$, where q_i represents player *i*'s beliefs about *q*. Let us denote this second threshold θ.

Figure 2: the trust game with an exogenous collective punishment mechanism.



In the following, we assume that $0 < \theta < \Theta < 1$, which is consistent with the parameters we adopt in the experiment. If the proportion p of *H*-types in society is larger than the threshold θ, and this is common knowledge, then in the game with collective punishment, any player *i* will never cheat as a second mover and will always trust as a first mover, regardless of his own type. If instead $p < \theta$, this game becomes a coordination game with two Pareto-ranked equilibria. In the low-efficiency equilibrium, *L*-type players cheat in the role of second mover, and nobody trusts as a first mover. In the high-efficiency equilibrium, instead, neither *L*-types nor *H*-types cheat as second movers, and everybody trusts as a first mover. There exists, however, the risk of miscoordination, as subjects cannot be certain of the strategy the others will adopt.

Let β_i be player *i*'s belief about the fraction of the other players who adopt the cooperative strategy *(trust, do not cheat)* in the trust game with a collective punishment mechanism. Then, we could obtain the belief q_i about the total number of players who will not cheat, which depends on two elements: the proportion of intrinsically trustworthy players p, and the belief β_i: $q_i = p + (1-p)\beta_i$

To summarize, for any value of p, the introduction of a collective punishment mechanism does not

decrease trustworthiness with respect to the baseline scenario, and might increase both trust and trustworthiness, if the proportion of *H*-types p is high enough, or if a sufficiently high number of players have high beliefs β_i about the fraction of the other players who adopt the cooperative strategy.[7]

Hypothesis 1: *In presence of a collective punishment mechanism, the levels of trust and trustworthiness are equal or higher than in the baseline scenario.*

3.2. Endogenous adoption of the collective punishment mechanism

We now consider the case in which, prior to playing the game (and before roles are assigned), players express their preference on whether to have or not a collective punishment mechanism in place. More specifically, we consider the case in which the implementation of the punishment mechanism is determined by a majority voting rule. The main question we would like to pursue is whether this mechanism can affect the beliefs q_i, thus serving as a coordination device to drive the society towards the efficient equilibrium.

Let us consider again the behavior of player i in the game with a collective punishment mechanism in place. Depending on the player *i*'s belief q_i, we can envisage five possible cases based on the types of players. For the *L*-type, i.e. selfish players, there are three possible scenarios:

(i.) $q_i \leq \theta < \Theta \leq 1$: the player chooses the strategy *(do not trust, cheat)*;
(ii.) $\theta < q_i < \Theta \leq 1$: the player chooses the strategy *(do not trust, do not cheat)*;
(iii.) $q_i \geq \Theta$: the player chooses the strategy *(trust, do not cheat)*.

For the *H*-type, i.e. intrinsically trustworthy players, there are two possible scenarios:

(iv.) $q_i < \Theta \leq 1$: the player chooses the strategy *(do not trust, do not cheat)*;
(v.) $q_i \geq \Theta$: the player chooses the strategy *(trust, do not cheat)*.

[7] An alternative, behavioral hypothesis, which we do not explore here, is that the exogenous introduction of a punishment mechanism could crowd out intrinsic motivations for trustworthiness (Bowles and Polania-Reyes, 2012; Fehr and Rockenbach, 2003).

However, these boil down to the first three scenarios, as (ii) and (iv) coincide, as well as (iii) and (v). Let us now calculate the player's expected profit in the trust game with collective punishment, under these three alternative scenarios. Remember that in the basic trust game, when $p < \Theta < 1$, player *i*'s expected profit is equal to x, no matter what, while if $p > \Theta$, then in the basic trust game player i would trust as a first mover, and everyone else does the same. In this case his expected payoff depends on his type.

Scenario (i). As a first mover, player i will not trust, hence he will be sure to earn x. As a second mover he will earn x if his partner does not trust, and $2x + 2s - zq_i$ if his partner chooses to trust. Because β_i is player *i*'s belief about the fraction of other players who adopt the cooperative strategy *(trust, do not cheat)*, he will expect the former event to take place with probability $1 - \beta_i$, and the latter with probability β_i. Hence, the expected profit a player can obtain in the game with collective punishment is:

$$E(\pi^s) = \frac{1}{2}x + \frac{1}{2}[x(1-\beta_i) + (2x + 2s - zq_i)\beta_i] = x + \frac{\beta_i}{2}(x + 2s - zq_i)$$

The expected profit above is greater than x if $q_i < \frac{x+2s}{z}$, which is true for every $q_i \leq \theta = \frac{x+s}{z}$. Hence, a selfish player with belief $q_i \leq \theta$ will prefer to have the punishment mechanism in place.

Scenario (ii). As a first mover, the player i will not trust, hence he will be sure to earn x. As a second mover he will earn x if his partner does not trust, which happens with probability $1 - \beta_i$, and $x + s$ if his partner chooses to trust, which happens with probability β_i. Hence, the expected profit in the game with collective punishment is:

$$E(\pi^s) = \frac{1}{2}x + \frac{1}{2}[x(1-\beta_i) + (x + s)\beta_i] = x + \frac{\beta_i}{2}s \geq x$$

Hence, both a selfish player and an intrinsically trustworthy player with beliefs $\theta < q_i < \Theta$ will prefer to have the collective punishment mechanism in place.

Scenario (iii). As a first mover, player i will trust, hence he will earn $x + s$ with probability q_i and 0 with probability $1 - q_i$. As a second mover he will earn x if his partner does not trust, which

happens with probability $1-\beta_i$, and $x+s$ if his partner trusts, which happens with probability β_i. Hence, the expected profit a player can obtain in the game with collective punishment is:

$$E(\pi^s) = \frac{1}{2}q_i(x+s) + \frac{1}{2}[x(1-\beta_i) + (x+s)\beta_i] = \frac{1}{2}[q_i(x+s) + x + \beta_i s]$$

In this case, however, the expected payoff $E(\pi^b)$ in the basic trust game depends on player i's type, and on whether $p > \Theta$. If $p < \Theta \leq q_i$ then $E(\pi^b) = x < E(\pi^s)$ and player i will vote in favor of the punishment mechanism. Indeed, the expected profit in presence of collective punishment is grater than x if $q_i(x+s) + \beta_i s > x$, which holds for every $q_i \geq \Theta = \frac{x}{x+s}$. Hence, both a selfish player and an intrinsically trustworthy player with $p < \Theta \leq q_i$ will prefer to have the punishment mechanism in place.

If instead $p > \Theta$, the preferences of *H*-type and *L*-type players will differ. If player i is an *H*-type, in the basic trust game as a first mover he will trust, hence expecting to earn $x+s$ with probability p and 0 with probability $1-p$. As a second mover he earns $x+s$ because all first movers should trust. Hence, the expected profit a player can obtain is:

$$E(\pi^b) = \frac{1}{2}p(x+s) + \frac{1}{2}(x+s) = \frac{1+p}{2}(x+s)$$

Consider also that if $p > \Theta$ then $q_i = \beta_i = 1$ for all players. Hence $E(\pi^s) = x + s \geq E(\pi^b)$: when $p > \Theta$, *H*-type players will always vote in favor of collective punishment.

By contrast, if player i is an *L*-type, in the basic trust game as a first mover he will trust, hence he will earn $x+s$ with probability p and 0 with probability $1-p$. As a second mover he earns $2(x+s)$ because all first movers should trust, and he will cheat. Hence, the expected profit a player can obtain is:

$$E(\pi^b) = \frac{1}{2}p(x+s) + (x+s) = \frac{2+p}{2}(x+s) > x + s = E(\pi^s)$$

Hence, when $p > \Theta$, *L*-type players will vote against collective punishment.

Hypothesis 2: *H-type players will always vote in favor of the introduction of a collective punishment mechanism; L-type players will also vote in favor of it, unless the proportion of H-types is sufficiently high to induce them to trust in the Baseline ($p > \Theta$).*

Consequently, the collective punishment mechanism will always be adopted if $\Theta \geq 0.5$, which is the case in our experiment. Hence, we can state the following hypothesis on the effects of the vote on trust and trustworthiness.

Hypothesis 3: *a majority vote in favor of the collective punishment mechanism does not reveal anything on the distribution of types and beliefs, hence it should not affect trust and trustworthiness levels, as compared to those observed when the mechanism is exogenously introduced.*

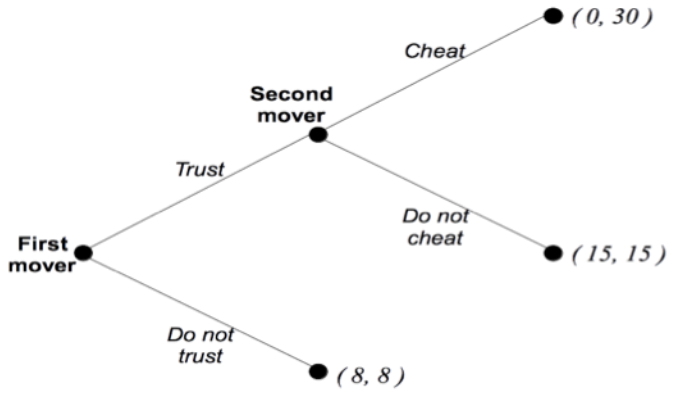
4. Experimental design

Our experimental treatments were based on variants of the binary-choice trust game (Bohnet et al., 2008) introduced in the previous section. We adopted a within-subject design, in which each participant was exposed to three treatments: Baseline, Exogenous and Voting. At the beginning of the session each subject was assigned to a group of six. In each treatment, subjects were paired with one of their group's members, to play a one-shot game. Matching across treatments was done so to ensure that no two subjects would meet more than once.[8] The group composition was kept constant during the whole session.

In the *Baseline* treatment, subjects were asked to play the binary trust game (i.e. Baseline game), as parameterized and represented in **Figure 3**. We adopted the strategy method (Brandts and Charness, 2011): all subjects had to choose their action both as a first mover and as a second mover, before knowing which role they would be assigned. Once all subjects had made their two choices, roles were randomly assigned, and subjects were matched in pairs. In each pair, payoffs were determined by the choice each of the two players had made for the role he was actually assigned.

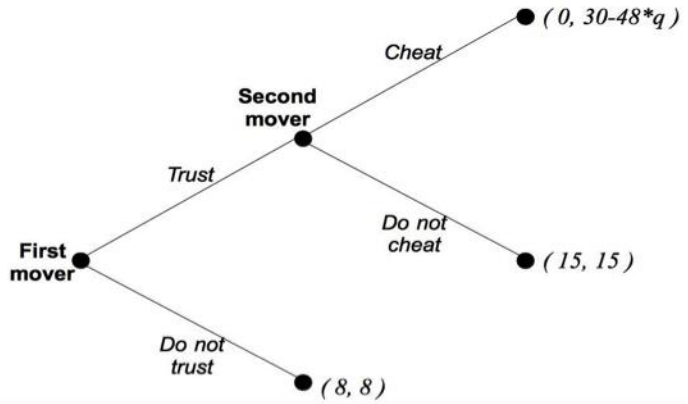
[8] With the exception of the Voting-IF treatment, as illustrated below.

Figure 3: Basic trust game -- parameterization adopted in the experiment.



In the *Exogenous* treatment, the strategic environment, the information structure and the options subjects had to choose were the same as in the *Baseline* game but, here, a collective punishment mechanism was exogenously introduced, under which cheating was sanctioned and the severity depended upon the number of subjects in the group, who chose not to cheat as second movers (i.e. Exogenous game, see **Figure 4**).[9]

Figure 4: Trust game with collective punishment -- parameterization adopted in the experiment.



After experiencing these two variations of the trust game, subjects entered the third treatment (*Voting*). At the beginning of this last treatment, before roles were assigned, subjects were asked to vote

[9] In order to be consistent with the theoretical model, in **Figure 4** the size of the sanction (*48*q*) is expressed in terms of the fraction *q* of subjects who choose not to cheat, in a group of six. In fact, in the experimental instructions, we expressed that variable as a function (*8*N*) of the number *N* of trustworthy players (see ***Appendix 2***). With the parameters adopted in our set up, we have that θ=0.3125 and Θ=0.5333. This implies that trusting is profitable even in the *Baseline* treatment, if the proportion of *H-types* in the society is higher than 0.5333, while if this proportion is as high as 0.3125, in the *Exogenous* treatment not cheating becomes more profitable than cheating.

for implementing either the *Baseline* or the *Exogenous* game, then a majority voting mechanism determined which of the two variations of trust games would have been played within the group, in this final phase. Abstention was not allowed. Before playing this third trust game, subjects were informed of the number of their group members who voted in favor of either option.

To reduce the risk of spillover effects, the outcomes of these three games were not revealed to the subjects, until the end of the session.[10] In addition, to control for possible order effects, in four sessions subjects were exposed to the *Baseline* treatment first, then they played the *Exogenous* treatment and finally the *Voting* treatment, while in other four sessions the order of the first two treatments was reversed.[11]

In order to examine whether having information about the aggregate behavior with and without collective punishment affected the individual voting behavior, in four of the sessions we introduced one additional treatment, after the *Voting* treatment. This treatment, denoted *Voting-IF*, was identical to the *Voting* treatment, with two exceptions. First, before voting subjects received information on the aggregate behavior of their group members in the Baseline and Exogenous treatments. More specifically, they were shown the number of subjects who chose either option, as a first and as a second mover, in each of the two treatments. Second, subjects were told that their partner might have been be the same person as in one of the previous three games.

Since our experiment was relatively complex, to ensure full understanding of the instructions, subjects were asked to complete a comprehension quiz with calculations and questions before making decisions in each stage game (see **Appendix 2**). Subjects were rewarded with €0.40 for each question they answered correctly at the first try. There were six questions per treatment (no questions before the *Voting-IF* treatment), hence subjects could earn up to €7.20 for the quiz.

At the end of the session, all subjects had to fill in a questionnaire including questions on their

[10] Each part of the instructions was distributed and read just before subjects started to play the corresponding game, which implies that subjects had no prior knowledge about the next part of the experiment. At the beginning of the experiment, subjects were simply informed that the experiment would be composed of three parts (see the Instructions in **Appendix 2**).
[11] For more information on the treatments and sessions, please refer to Table A in **Appendix 1**.

individual characteristics (gender, age, education, social status), general trust, risk attitudes, social preferences and cognitive abilities (see the **Appendix 3** for the complete text of the questionnaire). These questions allowed us to study how personal characteristics may affect the voting behavior, as well as the impact of the endogenous/exogenous introduction of collective punishment on individual behavior.

The experiment involved 96 subjects, divided in 8 sessions (see Table A in **Appendix 1**) and was conducted at the Bologna Laboratory for Experiments in Social Sciences (BLESS). Subjects were mostly undergraduate students at the University of Bologna, and were recruited through ORSEE (Greiner, 2015). About 53 percent of the subjects were male; nobody took part in more than one session. The experiment was programmed and implemented using the software z-Tree (Fischbacher, 2007). For each session, after showing up to the lab at the pre-scheduled session time, the 12 participants were randomly assigned to cubicles to avoid eye contact, and no communication was allowed during the experiment. The average session lasted about 75 minutes. Subjects were paid privately in cash at the end of the session and earned on average €18.25 (min. €9.5, max. €31.5) including the earnings from the comprehension quiz, which on average amounted to €6.4 (min. €3.2, max €7.2). No show-up fee was given.[12]

5. Results

In this section we carry out four steps of analysis. First, we juxtapose data from the *Baseline* and the *Exogenous* treatments, to analyze whether exogenously introducing collective punishment enhances the levels of trust and of trustworthiness in society. Second, we study subjects' voting behavior, and test whether most subjects vote in favor of collective punishment as predicted in our theoretical model. We also investigate who are the subjects who vote in favor of the punishment mechanism, and whether they differ from those who vote against it, along any significant dimension. Third, we examine whether the endogenous introduction of a collective punishment mechanism promotes efficiency by boosting trust and trustworthiness with respect to the case in which such a mechanism is exogenously imposed.

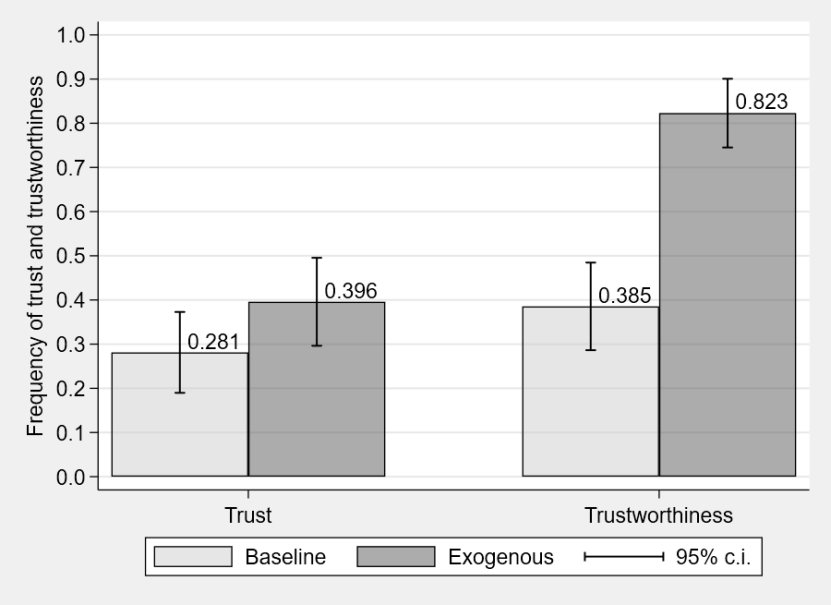
[12] For each session we recruited 15 subjects, to take into account possible no-show-ups, but only 12 students were randomly selected to participate. Supernumerary subjects were paid 5 Euros and had to leave before the session started.

We also study whether the endogenous choice not to adopt collective punishment depresses trust and trustworthiness, as predicted by our model. Finally, we examine whether the exposure to information about the aggregate behavior of their members in the *Baseline* and *Exogenous* treatments affects a subject's decision to vote in favor of the punishment mechanism.

5.1. Effects of collective punishment, when exogenously imposed

The main difference between *Baseline* and *Exogenous* games lies in the way the payoff of the player in the role of a second mover (i.e. trustee) depends on the other trustees' behavior, in case he chooses to abuse his partner's trust. This manipulation has a direct effect on trustworthiness and only an indirect effect on trust, because the player in the role of a first mover (i.e. trustor) will change her behavior only if she expects collective punishment to have a (direct) effect on the others' levels of trustworthiness. For this reason, we first present the results about trustees' behavior and then illustrate trustors' behavior.

Figure 5: frequency of trustful and trustworthy choices in the *Baseline* and *Exogenous* treatments.



Notes: One observation per subject, per treatment. The whiskers represent 95% confidence intervals.

As shown in **Figure 5**, the fraction of the trustworthy actions is larger when the collective punishment mechanism is exogenously imposed. More specifically, only 38.5% of subjects in the role

of trustee reciprocate trust in the *Baseline* treatment while 82.3% of trustees in the *Exogenous* treatment behave trustworthily. The difference is strongly significant ($p<0.001$). If not specified otherwise, comparisons across treatments are performed by means of logit regressions where the only explanatory variable is a treatment dummy.[13]

The impact of collective punishment on trustees' behavior emerges regardless of the order in which subjects are exposed to the *Baseline* and the *Exogenous* treatment, the level of trustworthiness being almost twice as high in the latter than in the former ($p<0.001$ in both cases, Table B in the **Appendix 1**). In addition, when we compare behavior across subjects, and focus exclusively on the first game played in each session, we observe that the difference in trustworthiness remains highly significant ($p<0.001$, Table B in the **Appendix 1**).

Figure 5 also shows that the overall level of trust is higher in the *Exogenous* than in the *Baseline* treatment. Specifically, while the average level of trust in the *Baseline* game is 28.1%, it reaches 39.6% in the *Exogenous* game, and the difference is marginally significant ($p=0.095$). However, if we control for the order effect, we find that when *Baseline* is implemented first the exogenously imposed punishment mechanism does not significantly enhance the trust ($p=0.837$). Conversely, when the punishment mechanism is implemented first but removed afterwards, the level of trust drops dramatically ($p=0.022$, see Table C in **Appendix 1**). We can summarize our results as follows.

Result 1: *the presence of a collective punishment mechanism significantly increases trustworthiness, and to a lesser extent also trust.*

5.2. Endogenous adoption of collective punishment

Our theoretical model predicts that, in the *Voting* treatment, *H*-types would always vote in favor of the collective punishment mechanism, while L-types would vote against it only if the proportion of H-types in society is very high (Hypothesis 2). Our data reveal instead that only a minority of subjects (30.2%) vote in favor of the mechanism, and that subjects' voting behavior does not seem to depend on

[13] Two-tailed z-tests using the pair of decisions made by each subject as an independent observation always confirm the results, both qualitatively and quantitatively.

their preferences or beliefs. This result does not depend on the order of the first two treatments: 29.2% of subjects vote in favor of the mechanism when the *Baseline* treatment is first played, while 31.2% opt for the punishment mechanism when subjects are first exposed to the *Exogenous* treatment, and the difference is not statistically significant (p=0.824).

Table 1: subjects' behavior in the *Baseline* treatment.

	Trust in *Baseline*		
Reciprocate in *Baseline*	Yes	No	Total
Yes (*H-type)*	20.8%	17.7%	38.5%
No (*L-type)*	7.3%	54.2%	61.5%
Total	28.1%	71.9%	100.0%

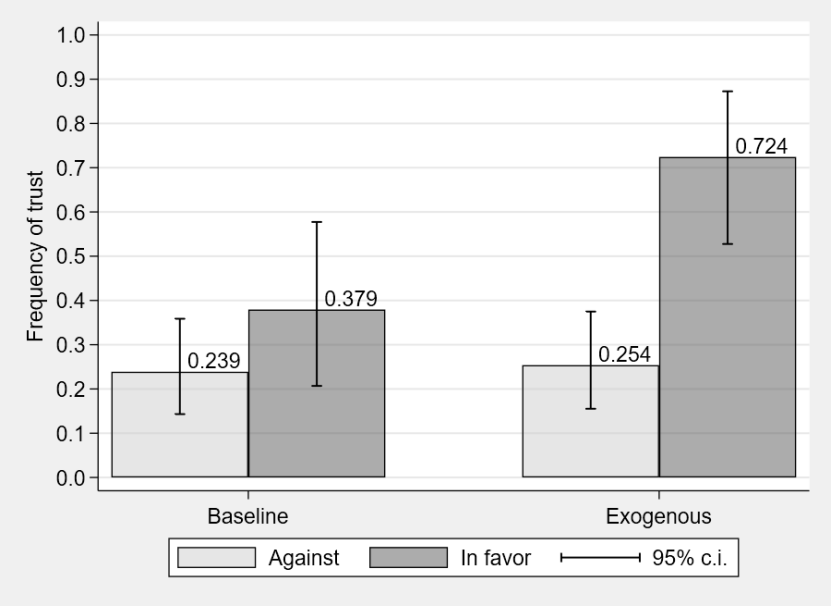
Since we adopt the strategy method in the experiment, for every subject we observe both choices (as a trustor and a trustee) in each treatment. We use subjects' behavior in the *Baseline* treatment to infer their preferences and beliefs. More specifically, first-movers' choices are a proxy for beliefs, since, as explained in Section 3, a subject in this treatment should trust only if he expects *H*-types to represent a high enough fraction of the population. We then classify subjects as *L*-types and *H*-types based on whether they cheat or not as trustees. Indeed, according to our model, *L*-type players would vote against the introduction of the punishment mechanism only if they trust in the *Baseline*. **Table 1** reports the distribution of subjects, along these two dimensions. It reveals that, according to our predictions, only 7.3% of the subjects should vote against the adoption of collective punishment, in the *Voting* treatment, while in our experiment this proportion was much higher.

To better understand the source of this discrepancy between our results and the theoretical predictions, we now investigate the determinants of subjects' voting decision. First, we divide subjects into two categories, depending on their voting decisions: against collective punishment and pro-punishment. We find that these two categories of subjects have similar levels of trust and trustworthiness in the *Baseline* treatment, implying that there is no difference in the preferences or ex-ante beliefs between them (p=0.165 for the difference in trust level, and p=0.409 for the difference

in trustworthiness level).[14]

In the *Exogenous* treatment, as revealed in **Figure 6**, subjects who vote in favor of collective punishment are more likely to trust their partners than others (72.4% vs. 25.4%, p<0.001). We also find that these pro-punishment subjects react more to the introduction of the punishment mechanism, i.e. they are more likely to increase their level of trust from the *Baseline* to the *Exogenous* game, as compared to the subjects who voted against the mechanism (p=0.002).

Figure 6: trust displayed by subjects who voted in favor and against collective punishment.



Notes: One observation per subject, per treatment. The whiskers represent 95% confidence intervals.

To dig deeper into these differences, we run three logit regressions (**Table 2**). The dependent variable indicates whether the subject voted in favor of collective punishment. In Model 1 we introduce subjects' choices in the *Baseline* as explanatory variables, finding that subjects' preferences and their ex-ante beliefs about others do not affect their voting behavior. In Model 2, we instead use their choices in the *Exogenous* treatment as explanatory variables. Our result shows that the probability that a subject votes in favor of collective punishment is 52.6% higher when she chose to trust and to reciprocate in

[14] P-values are obtained by means of logit regressions where the only explanatory variable is a dummy taking value one for subjects who voted in favor of the punishment mechanism in the Voting treatment. Results are confirmed by two-tailed z-tests using the decision made by each subject as an independent observation.

the *Exogenous* treatment. This strongly significant difference reappears in the Model 3, which includes the choices of subjects in both *Baseline* and *Exogenous*, suggesting that only those who interiorize the impact of collective punishment on trustworthiness and react to it with a higher level of trust, are inclined to vote in favor of it.

Result 2. *Only about 30% of subjects vote in favor of the collective punishment mechanism, and the voting behavior does not depend on subjects' preferences and beliefs.*

Table 2: Logit regressions on the determinants of subjects' voting behavior.

Dependent variable: Vote	**Model 1**	**Model 2**	**Model 3**
Choices in Baseline			
Trust & not Reciprocate	0.016		-0.149
	(0.181)		(0.111)
Not Trust & Reciprocate	-0.034		0.047
	(0.120)		(0.122)
Trust & Reciprocate	0.181		0.070
	(0.127)		(0.109)
Choices in Exogenous			
Trust & not Reciprocate		0.242	0.305
		(0.211)	(0.229)
Not Trust & Reciprocate		0.058	0.055
		(0.101)	(0.098)
Trust & Reciprocate		0.503***	0.526***
		(0.123)	(0.123)
Number of Observations	96	96	96

Notes: Marginal effects from logit regressions, with standard errors in parentheses. Trust-BL (Trustworthiness-BL) equals 1 for subjects choosing to trust (reciprocate) in the *Baseline* treatment; Trust-EX (Trustworthiness-EX) equals 1 for subjects choosing to trust (reciprocate) in the *Exogenous* treatment. The symbols *, **, and *** indicate significance a the 10%, 5% and 1% level, respectively.

Our next step is to explore the question of whether subjects' individual characteristics affect their

voting behavior. **Table 3** reveals that subjects who vote in favor of the punishment mechanism have higher cognitive abilities than the others, as supported by an ordered logit regression on the number of correct answers given to the three questions of the Cognitive Reflection Test. The result is confirmed if we look at the IQ test to measure subjects' cognitive abilities, which also reveals that subjects who vote in favor of collective punishment are significantly more likely to answer correctly.

Table 3: Individual characteristics and voting.

Individual characteristics	**Against (N=67)**	**In favor (N=29)**	**Significance of the difference**
Male	49.3%	62.1%	$p>0.1$[a]
Age	25.6	24.1	$p=0.065$[b]
CRT	1.1	1.6	$p=0.069$[c]
IQ	1.2	1.7	$p=0.003$[c]
Higher education	67.2%	44.8%	$p=0.042$[a]
Economics	50.7%	48.3%	$p>0.1$[a]
Statistics	44.8%	58.6%	$p>0.1$[a]
Game theory	28.4%	20.7%	$p>0.1$[a]
Trust	17.9%	17.2%	$p>0.1$[a]
Altruism	7.8	8	$p>0.1$[c]
Risk aversion	5.8	5.2	$p>0.1$[c]
RightAnswerBL	5.4	5.2	$p>0.1$[c]
RightAnswerEXO	5.1	5.2	$p>0.1$[c]
RightAnswerVOTE	5.6	5.6	$p>0.1$[c]

Notes: *Male* is a dummy taking value 1 for males and 0 for females; *Age* indicates subjects' age; *Higher education* equals 1 for those who have obtained at least a bachelor degree, and 0 otherwise; *CRT* ranges between 0 and 3 and is calculated by a three-item cognitive reflection test introduced by Frederick (2005); *IQ* ranges between 0 and 2 and is calculated by a two-item IQ test; *Economics*, *Statistics*, and *Game theory* are dummies taking value 1 for those who have taken at least one course in economics, statistics, or game theory, respectively; *Trust* equals 1 for those whose answer to the WVS on generalized trust is positive, and 0 otherwise; *Altruism* corresponds to our questionnaire-based measure of altruism; *Risk aversion* indicates subjects' answer to the risk attitude question; *RightAnswerBL*, *RightAnswerEXO*, and *RightAnswerVOTE* indicate the number of the correct answers to the control questions in the *Baseline, Exogenous*, and *Voting* treatment, respectively.
The symbols *, **, and *** indicate significance at the 10%, 5% and 1% level, respectively.
[a] Linear regression.
[b] Logit regression.
[c] Ordered logit regression.

Results in **Table 3** also indicate that, although our experimental design is relatively complicated, subjects could answer most of the control questions correctly before playing the game and, on average, those who voted against or in favor of the punishment mechanism could provide a similar number of right answers. This implies that all subjects could well understand the instructions, and that differences

in the voting behavior are not driven by comprehension problems.

Table 4: Voting behavior and individual characteristics

Dependent variable: Vote	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Male	0.125					0.022
	(0.091)					(0.097)
Age	-0.032**					-0.019
	(0.016)					(0.016)
Higher education		-0.218**				-0.095
		(0.097)				(0.105)
Economics		0.008				-0.010
		(0.096)				(0.091)
Statistics		0.205**				0.182*
		(0.095)				(0.094)
Game theory		-0.131				-0.099
		(0.107)				(0.113)
CRT			0.016			0.034
			(0.045)			(0.049)
IQ			0.227***			0.230**
			(0.081)			(0.095)
Trust				0.003		0.021
				(0.123)		(0.114)
Altruism				0.025		0.011
				(0.029)		(0.030)
Risk aversion				0.036		0.040*
				(0.023)		(0.022)
RightAnswerBL					-0.051	-0.054
					(0.053)	(0.053)
RightAnswerEXO					0.023	-0.014
					(0.056)	(0.051)
RightAnswerVOTE					0.020	-0.057
					(0.079)	(0.074)
N. Observations	96	96	96	96	96	96

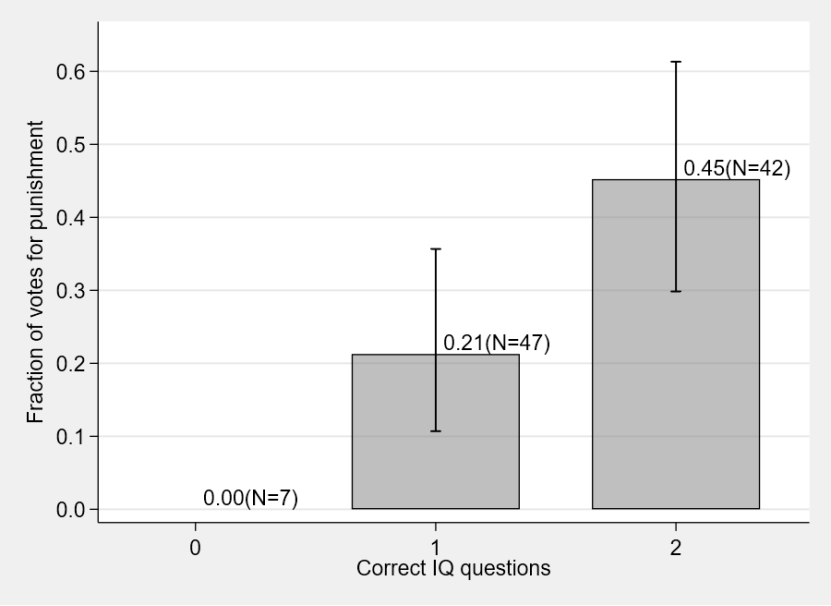
Notes: Marginal effects from logit regressions (standard errors reported in parentheses). *CRT* ranges between 0 and 3 and is calculated by a three-item cognitive reflection test introduced by Frederick (2005); *IQ* ranges between 0 and 2 and is calculated by a two-item IQ test. *Economics*, *Statistics*, and *Game theory* are dummies taking value 1 for those who have taken at least one course in economics, statistics, or game theory, respectively.

The symbols *, **, and *** indicate significance at the 10%, 5% and 1% level, respectively.

Table 4 reports results from a set of logit regressions providing further support for this result. The dependent variable is a dummy taking value one for the subjects who voted in favor of collective

punishment. In Model 1 we control for subjects' demographics, which do not correlate with their voting behavior. In Model 2 we include as regressors three dummy variables meant to capture the academic background of the subjects. Results indicate that subjects who have some prior knowledge of statistics are more likely to vote in favor of the punishment mechanism. Model 3, where the only explanatory variables are *IQ* and *CRT*, indicates that cognitive abilities measured by the Cognitive Reflection Test are not significantly correlated with subjects' voting behavior, while the probability of voting for the punishment mechanism is 22.7% larger among subjects who were able to answer one more of the two IQ questions. Model 4 shows that individual preferences, as measured by our post-experimental questionnaire, do not correlate with voting either, and Model 5 confirms that voting is not affected by subjects' understanding of instructions. Model 6 includes all regressors and confirms that cognitive abilities seem to be the main driver of voting behavior: subjects need to be sophisticated enough to fully anticipate the consequences of the introduction of collective punishment, hence its profitability. This is also illustrated in **Figure 7**.

Figure 7: IQ and voting.



Notes: One observation per subject, per treatment. The whiskers represent 95% confidence intervals.

5.3. Effects of the endogenous adoption or rejection of the punishment mechanism

The experimental literature on public good games has shown that there is a "dividend of democracy" in the sense that institutions endogenously chosen through voting can be more efficient than the same institutions being exogenously imposed on decision makers (Dal Bo et al., 2010; Sutter et al., 2010). One possible reason is that voting for the deterrent (or non-deterrent) institutions that punish uncooperative subjects credibly signals an intention to establish a high level of cooperation and thereby induces other group members to do the same. Consequently, the voting mechanism promotes coordination on the efficient, cooperative outcome (Markussen et al., 2014). Here we investigate whether "the dividend of democracy" can be observed in our setting. Specifically, we investigate whether the punishment mechanism, when endogenously chosen, significantly increases the levels of trust and trustworthiness relative to the case in which it is exogenously imposed.

In our study, only three groups endogenously adopt the collective punishment mechanism, while the other thirteen groups play the baseline trust game in the Voting treatment. Consider the behavior of subjects in the role of trustee first. When most of the group members vote against the implementation of collective punishment, the average level of trustworthiness does not change substantially, decreasing from 34.6% to 33.3%, relative to the *Baseline* treatment. Similarly, in groups where collective punishment is endogenously adopted trustworthiness levels decreased from 100% to 94.4%, relative to the *Exogenous* treatment. Neither difference is statistically significant (two-sided z-tests: p=0.866, N=78 for the former comparison, and p=0.311, N=18 for the latter). Similar results emerge if we focus on trust: when subjects vote for not introducing the punishment mechanism the level of trust drops from 25.6% to 23.1%, compared to the *Baseline*; the fraction of trustful behavior is 55.6% -- exactly as in the *Exogenous* -- in groups where the collective punishment mechanism is determined by the majority voting mechanism. These two differences are also not statistically significant (two-sided z-tests: p=0.709, N=78 for the former, and p=1.000, N=18 for the latter).

Result 3. *When subjects vote for (not) introducing collective punishment, the levels of trust and trustworthiness are not significantly different from the case in which collective punishment is exogenously (not) introduced.*

While "the dividend of democracy" has been often observed in previous experimental papers, our

study fails to find any positive effect of the voting mechanism on the society's ability to coordinate on an efficient outcome. A possible reason is that, when the mechanism is endogenously chosen, only those who voted in favor of it react positively. Indeed, we find that the three groups where the punishment mechanism was endogenously activated achieved a higher level of trustworthiness in the *Exogenous* treatment: the average level of trustworthiness is 100% in these groups and 78.2% in the other thirteen, and the difference is significant (two-sided z-test: N1=78, N2=18, p=0.029). These three groups also exhibit higher levels of trust than other groups in the *Exogenous* treatment. The average level of trust is 55.6% in the three groups where the punishment mechanism is endogenously imposed and 35.9% in the other thirteen groups, but the difference is not significant (two-sided z-test: N1=78, N2=18, p=0.124).

In addition, within these three groups, not all subjects positively react to the collectively determined punishment mechanism. Our results suggest that the endogenously chosen mechanism induces only those who vote in favor of it to be more trustful, while if anything the others trust less. When collective punishment is endogenously chosen, the 10 subjects who vote in favor of it increase their trust level from 70% to 80% as compared to the *Exogenous* treatment, while other 8 subjects reduce their trust level from 37.5% to 25%. Due to the limited sample, however, we cannot detect whether these differences are statistically significant.

5.4. Effects of information about others' behavior on voting

We now turn to the question of whether feedback about the aggregate behavior in the group, with and without collective punishment, could help subjects understand the effectiveness of the punishment mechanism, thereby changing their voting behavior. In the last 4 experimental sessions, we added a fourth game, where subjects received information on the aggregate behavior of their group members in the *Baseline* and *Exogenous* treatments before deciding whether to vote for or against collective punishment (see Section 3).

Among the 48 subjects who took part in these additional sessions, only 8 (i.e. 16.7%) changed their vote after observing the aggregate information about the first two treatments. Of them, five subjects

voted in favor of collective punishment in the *Voting-IF* treatment, and three voted against it. A two-sided z-test indicates that there is no difference in the voting behavior between in the *Voting* and *Voting-IF* treatments (N=48, p=0.653). Only two groups endogenously adopted the collective punishment mechanism in the last treatment. To explore subjects' voting behavior in more depth, we run two logit regressions, whose results are reported in **Table 5**.

Table 5: Voting behavior and feedback information.

Dependent variable: Vote	**Model 1**		**Model 2**	
Diff-Trust	0.009	(0.039)	-0.268*	(0.152)
Diff-Trustworthiness	0.061	(0.064)	0.249	(0.210)
IQ	0.214*	(0.112)	0.504**	(0.239)
IQ x Diff-Trust			0.194***	(0.074)
IQ x Diff-Trustworthiness			-0.135	(0.104)
N. Observations	48		48	

Notes: Marginal effects from logit regressions (standard errors robust for clustering at the matching-group level are reported in parentheses). The symbols *, **, and *** indicate significance at the 10%, 5% and 1% level, respectively.

The dependent variable is a dummy taking value one when the subject voted in favor of collective punishment. *Diff-Trust (Diff-Trustworthiness)* indicates the difference between the number of the other group members who are trustful (trustworthy) in the *Exogenous vs. Baseline* treatments; *IQ* ranges between 0 and 2 and is calculated by a two-item IQ test. Model 1 shows that on average subjects do not react to information on the effect that the punishment mechanism has on trust and trustworthiness. Model 2 reveals that in fact only subjects with better cognitive abilities took this information into account in the *Voting-IF* treatment.

Result 4. *Even with information about others' past behavior, most subjects do not change their vote.*

6. Discussion and Conclusions

In this paper, we explore whether the endogenous adoption of a collective punishment mechanism can help a society coordinate on an efficient outcome, characterized by high levels of trust and trustworthiness. We first introduce a theoretical analysis of the consequences of the introduction of a

collective punishment mechanism, which largely builds upon Anderlini and Terlizzese's (2017) work. We then design and run an experiment to empirically test our theoretical predictions.

We find that subjects exhibit significantly higher levels of trust and trustworthiness when a collective punishment mechanism is imposed exogenously. In contrast with the previous studies on the "dividend of democracy", however, we fail to observe that the punishment mechanism induces higher level of cooperation when it is democratically chosen compared to the case in which it is exogenously activated. One potential explanation is that in most previous studies subjects could directly inflict punishment on low contributors to the public good to enforce the endogenously determined rule, or the punishment was fixed and determined ex-ante by the experimenter. By contrast, in our trust game, even when the social sanction is democratically introduced, the severity of punishment depends on the average behavior in society, which makes it more unpredictable from the subjects' perspective; hence a higher cognitive effort is necessary to anticipate how others will react to the rule, and to predict its overall effects on profits and welfare. Further experimental studies are needed to more precisely pin down the mechanisms driving these differences.

Another important finding is that most subjects vote against the collective punishment mechanism, even though from an ex post perspective it would have paid off, on average, to vote in favor of it. Previous experimental studies have shown that subjects are reluctant to choose a punishment institution when facing alternative options. In Sutter et al. (2010), subjects are allowed to vote for a voluntary contribution mechanism (VCM), an institution with reward possibility and an institution with punishment possibility. The authors report that under unanimous voting, the punishment option is rarely selected. A similar behavior pattern is also observed in Botelho et al. (2007). After having experienced both the VCM and the VCM with the punishment option, subjects decide to choose the governing institution for the final period. Botelho et al. (2007) find that in their experiment 77.8% of subjects vote against the punishment institution. One possible reason is that subjects may naturally dislike the punishment since it evokes negative feelings. To test whether opting against the sanction is mainly driven by a "natural aversion" to punishment, in future research we plan to run a follow-up experiment where we reframe the game without changing the incentives, and substitute penalties with

rewards. Another potential explanation is that cognitive limitations may refrain subjects from anticipating the positive effect of the introduction of collective punishment. Putterman et al. (2011) find that intelligence predicts subjects' votes on efficient schemes when they are permitted to vote over a menu of sanction rules. Our study also confirms that subjects with high cognitive abilities are more likely to anticipate the effectiveness of collective punishment and therefore vote in favor of it.

In an additional treatment, we investigate whether the information about the others' aggregate behavior with and without collective punishment affects subjects' voting choices, finding that subjects hardly change their votes respect to the no-feedback condition. In Gurerk (2013), before a voting phase in which they choose among alternative institutions governing the public good provision, subjects are provided with the complete history of a punishment institution which was actually implemented in a previous experiment. The author finds that social information significantly induces more subjects to accept the punishment option and reach full contributions more quickly over time. Our study fails to replicate the positive effect of social information, a result which is in line with some previous studies, showing that a high percentage of subjects are reluctant to select a relatively efficient mechanism even when they are exposed to the complete information on subjects' behavior under the alternative institutional regimes (Dal Bo et al., 2010; Gurerk, et al., 2006; Hilbe, et al., 2014). One possible reason is that subjects may need repetition to fully understand the change in incentives introduced by the collective punishment mechanism, and its effects on others' behavior; we see this as an interesting route for future research. Another possible way of promoting the endogenous adoption of an efficiency-enhancing institution is group communication. Alm et al. (1999) investigate the effect of voting on a social norm of tax compliance by letting subjects vote via majority rule on different aspects of the fiscal system. They find that, without communication, subjects vote against an increase in the levels of sanction enforcement imposed on tax evaders. However, when subjects are allowed to communicate before voting, they are more likely to select a greater level of enforcement, achieving an overall increase in efficiency. Along these lines, we could also expand our set-up and examine the question of whether group communication before the voting phase facilitates the acceptance of the collective punishment institution. All this, however, is left for future research.

References

Algan, Y., and Cahuc, P., 2013, "Trust and Human Development: Overview and Policy", *Handbook of Economic Growt*h, ed. by Philippe Aghion and Steven Durlauf.

Alm, J., McClelland, G., and Schulze, W., 1999, "Changing the Social Norm of Tax Compliance by Voting." *Kyklos*, 52, 141-171.

Anderlini, L. and Terlizzese, D., 2017, "Equilibrium Trust", *Games and Economic Behavior*, 102, 624-644.

Andreoni, J., and Gee, L. K., 2012, "Gun for hire: delegated enforcement and peer punishment in public goods provision." *Journal of Public Economics* 96, 1036-1046.

Becker, G. S., 1968, "Crime and Punishment: An Economic Approach", *Journal of Political Economy*, 76, 169-217.

Bernstein, L., 1992, "Opting out of the Legal System: Extralegal Contractual Relations in the Diamond Industry", *Journal of Legal Studies*, 21, 115–157.

Bernstein, L., 2001, "Private Commercial Law in the Cotton Industry: Creating Cooperation through Rules, Norms, and Institutions", *Michigan Law Review*, 99, 1724–90.

Besley, T., and Coate, A., 1995, "Group Lending, Repayment Incentives and Social Collateral." *Journal of Development Economics*, 46(1), 1-18.

Blanco, M., Engelmann, D., and Normann, H. T., 2011, "A Within-subject Analysis of Other-regarding Preferences." *Games and Economic Behavior*, 72, 321-338.

Bohnet, I., and Cooter, R., 2003, "Expressive Law: Framing or Equilibrium Selection?" KSG Working Paper No. RWP03-046; and UC Berkeley Public Law Research Paper No. 138. http://ssrn.com/abstract=452420.

Bohnet, I., Grieg, F., Herrmann, B., and Zeckhauser, R., 2008, "Betrayal Aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States." *American Economic Review*, 98(1), 294– 310.

Botelho, A., Harrison, G., Costa Pinto, L., and Rutstrom, E., 2007, "Social Norms and Social Choice." Working paper 05-23, Department of Economics, College of Business Administration, University of Central Florida.

Bowles, S., and Polania-Reyes, S., 2012, "Economic Incentives and Social Preferences: Substitutes or Complements?" *Journal of Economic Literature*, 50, 368-425.

Brandts, J., and Charness, G., 2011, "The Strategy versus the Direct-Response Method: A First Survey of Experimental Comparisons", *Experimental Economics*, 21, 1-24.

Buchan, N. R., Croson, R. T., & Dawes, R. M., 2002, "Swift Neighbors and Persistent Strangers: A Cross-Cultural Investigation of Trust and Reciprocity in Social Exchange." *American Journal of Sociology*, 108(1), 168-206.

Charness, G., Cobo-Reyes, R., and Jimenez, N., 2008, "An Investment Game with Third-Party Intervention." *Journal of Economic Behavior & Organization*, 68, 18-28.

Cialdini, R., Reno, R., and Kallgren, 1990, "A Focus Theory of Normative Conduct: Recycling the Concept of Norms to Reduce Littering in Public Places." *Journal of Personality and Social Psychology*, 58, 1015-1026.

Clay, K., 1997, "Trade without Law: Private-order Institutions in Mexican California", *Journal of Law, Economics, and Organization*, 13, 202–231.

Clay, K., and Wright, G., 2005, "Order without Law? Property Rights during the California Gold Rush", *Explorations in Economic History*, 42, 155-183.

Cooter, R., 1998, "Expressive Law and Economics", *Journal of Legal Studies*, 27, 585–608.

Dal Bó, E., Dal Bó, P., & Eyster, E., 2017, "The Demand for Bad Policy when Voters Underappreciate Equilibrium Effects." *The Review of Economic Studies.*

Dal Bo, P., Foster, A., and Putterman, L., 2010, "Institutions and Behavior: Experimental Evidence on the Effects of Democracy", *American Economic Review*, 100, 2205-2229.

de Quervain, D., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., and Fehr, E., 2004, "The Neural Basis of Altruistic Punishment." *Science*, 305, 1254-1258.

Ertan, A., Page, T., and Putterman, L., 2009, "Who to Punish? Individual Decisions and Majority Rule in Mitigate the Free Rider Problem." *European Economic Review*, 53, 495-511.

Fehr, E., and Williams, T., 2013, "Endogenous Emergence of Institutions to Sustain Cooperation." Working Paper, University of Zurich.

Fehr, E., and Rockenbach, B., 2003, "Detrimental Effects of Sanctions on Human Altruism." *Nature*, 422, 137-140.

Fischbacher, U., 2007, "z-Tree: Zurich Toolbox for Ready-made Economic Experiments." *Experimental Economics*, 10, 171-178.

Frederick, S., 2005, "Cognitive Reflection and Decision Making." *Journal of Economic Perspective*, 19(4), 25-42.

Galbiati, R., and Vertova, P., 2008, "Obligations and Cooperative Behavior in Public Good Games." *Games and Economic Behavior*, 146-170.

Greif, A., 1989, "Reputation and Coalitions in Medieval Trade: Evidence on the Maghribi Traders", *Journal of Economic History*, XLIX, 857–82.

Greif, A., 1993, "Contract Enforceability and Economic Institutions in Early Trade: the Maghribi Traders' Coalition", *American Economic Review*, 83, 525–48.

Greif, A., 2006, *Institutions and the Path to the Modern Economy: Lessons from Medieval Trade*, Cambridge University Press.

Greiner, B., 2015, "Subject Pool Recruitment Procedures: Organizing Experiments with ORSEE." Journal of the Economic Science Association 1 (1), 114-125.

Guiso, L., Sapienza, P., and Zingales, L., 2008, "Social Capital as Good Culture." *Journal of the European Economic Association*, 6(2-3), 295-320.

Gurerk, O., 2013, "Social Learning Increases the Acceptance and the Efficiency of Punishment Institutions in Social Dilemmas." *Journal of Economic Psychology*, 34, 229-239.

Gurerk, O., Irlenbusch, B. and Rockenbach, B., 2006, "The Competitive Advantage of Sanctioning Institutions." *Science*, 312, 108–111.

Gurerk, O., Irlenbusch, B. and Rockenbach, B., 2014, "On Cooperation in Open Communities." *Journal of Public Economics*, 120, 220-230.

Hilbe, C., Traulsen, A., Rohl, T., and Milinshi, M., 2014, "Democratic Decisions Establish Stable Authorities That Overcome the Paradox of Second-order Punishment." *Proceedings of the National Academy of Sciences*, 111, 752-756.

Huang, P., and Wu, H., 1994, "More Order without More Law: A Theory of Social Norms and Organizational Cultures." *Journal of Law, Economics, and Organization*, 10(2), 390-406.

Kadens, E., and Young, E., 2013, "How Customary Is Customary International Law?" *William & Mary Law Review*, 54, 885-920.

Kamei, K., 2014, "Democracy and Resilient Pro-Social Behavioral Change: An Experimental Study", Available at SSRN: http://dx.doi.org/10.2139/ssrn.1756225.

Kimbrough, E., and Rubin, J., 2015, "Sustaining Group Reputation." *Journal of Law, Economics, & Organization*, 31, 599-628.

Kosfeld, M., Okada, A., and Riedl, A., 2009, "Institution Formation in Public Good Games." *American Economic Review*, 1335-1355.

Leeson, P., and Williamson, C., 2009, "Anarchy and Development: An Application of the Theory of Second Best." *Law and Development Review*, 2(1), 76-96.

Markussen, T., Putterman, L., and Tyran, J., 2014, "Self-Organization for Collective Action: An Experimental Study of Voting on Sanction Regimes." *Review of Economic Studies*, 81, 301-324.

Masclet, D., Noussair, C., Tucker, S.L., Villeval, M.C., 2003, "Monetary, very weakness and nonmonetary punishment in the voluntary contributions mechanism." *The American Economic Review*, 93, 366-380.

McAdams, R., 2000a, "An Attitudinal Theory of Expressive Law." *Oregon Law Review*, 79, 339–390.

McAdams, R., 2000b, "A Focal Point Theory of Expressive Law." *Virginia Law Review*, 86, 1649–1731.

McAdams, R., and Nadler, J., 2005, "Testing the Focal Point Theory of Legal Compliance: The Effect of Third-Party Expression in an Experimental Hawk/Dove Game." *Journal of Empirical Legal Studies*, 2(1), 87-123.

Ostrom, E., 1990, *Governing the Commons: The Evolution of Institutions for Collective Action*, Cambridge University Press, New York.

Polinsky, A. M. and Shavell, S., 2000, "The Economic Theory of Public Enforcement of Law." *Journal of Economic Literature*, 38, 45-76.

Polinsky, A. M. and Shavell, S., 2008, "Economic Analysis of Law", *The New Palgrave Dictionary of Economics*, ed. by Steven N. Durlauf and Lawrence E. Blume.

Posner, E., 1998, "Symbols, Signals, and Social Norms in Politics and the Law." *Journal of Legal Studies*, 27, 765–789.

Posner, E., 2000, *Law and Social Norms*, Harvard University Press, Cambridge, MA.

Posner, R., 1997, "Social Norms and the Law: An Economic Approach." *American Economic Review*, 87(2), 365-369.

Posner, R., and Rasmusen, E., 1999, "Creating and Enforcing Norms, with Special Reference to Sanctions." *International Review of Law and Economics*, 19, 369-382.

Putterman, L., Tyran, J., and Kamei, K., 2011, "Public Goods and Voting on Formal Sanction Schemes." *Journal of Public Economics*, 95, 1213-1222.

Samek Savikhin A., and Sheremeta, R.M., 2014, "Recognizing contributors: an experiment on public goods." *Experimental Economics* 17(4), 673-690.

Schultz, P. W., Nolan, J. M., Cialdini, R., B., Goldstein, N. J., and Griskevicius, V., 2007, "The Constructive, Destructive, and Reconstructive Power of Social Norms." *Psychological Science*, 18(5), 429-434.

Sutter, M., Haigner, S., and Kocher, M., 2010, "Choosing the Carrot or the Stick? Endogenous Institutional Choice in Social Dilemma Situations." *Review of Economic Studies*, 77, 1540-1566.

Tabellini, G., 2008, "The Scope of Cooperation: Norms and Incentives." *Quarterly Journal of Economics*, 123(3), 905-950.

Tyler, T., 1990, *Why People Obey Law*, Yale University Press.

Tyran, J., and Feld, L., 2006, "Achieving Compliance when Legal Sanctions are Non-deterrent." *Scandinavian Journal of Economics*, 108(1), 135-156.

Vollan, B., 2011, "The Difference between Kinship and Friendship: (Field-) Experimental Evidence on Trust and Punishment." *The Journal of Socio-Economics*, 40, 14-25.

Vollan, B., Zhou, Y., Landmann, A., Hu, B., Herrmann-Pillath, C., 2013, "Cooperation and Authoritarian Norms: An Experimental Study in China." Working Papers in Economics and Statistics, University of Innsbruck.

Xiao, E., 2013, "Profit –Seeking Punishment Corrupts Norm Obedience." *Games and Economic Behavior*, 77, 321-344.

Authors

Huojun Sun was a research affiliate at the Department of Economics of the University of Bologna. He obtained a Ph.D. in Law and Economics in 2015, from a joint program involving the Hamburg Institute of Law & Economics, the Department of Economics of the University of Bologna, and the Rotterdam Institute for Law & Economics. His main research interest focused on interpersonal trust, social norms and social conflicts using experimental economics methods. Huojun Sun prematurely died on January 11, 2018, after a 2-year long battle against cancer.

Maria Bigoni is Associate Professor at the Department of Economics of the University of Bologna. She has published in international journals such as Econometrica, the RAND Journal of Economics, the Economic Journal, and Games and Economic Behavior. Her main research interest is experimental economics, applied to the study of cooperation in repeated social dilemmas, industrial organization and learning. Her most recent line of research focuses on the effects of economic inequality of cooperation.

Appendix 1

Table A: Treatments and sessions

Session type	**Baseline-first**	**Exogenous-first**	**Baseline-first +Information**	**Exogenous first + Information**
Order	BL-EX-VT	EX-BL-VT	BL-EX-VT-VF	EX-BL-VT-VF
Session dates	Dec. 03, 2013; Dec. 10, 2013	Dec.12, 2013	March 20, 2014 March 24, 2014	March 20, 2014 March 24, 2014
N. Subjects	24	24	24	24
N. Independent observations	24	24	24	24

Notes: In the table, BL stands for *Baseline*, EX for *Exogenous*, VT for *Voting*, and VF for *Voting-IF*.

Table B: Order effects on trustworthiness

	Trustworthiness (%)		
	1st Game		**2nd Game**
BL-EXO-VOTE	43.8%	$<^{***}$	77.1%
	$\wedge^{***}$		$\vee^{***}$
EXO-BL-VOTE	87.5%	$>^{***}$	33.3%

Notes: BL stands for *Baseline* treatment, EXO for *Exogenous* treatment, VOTE for *Voting* treatment. *** indicates the significance at 1% level based on logit regressions. The results are confirmed using two-tailed z-tests.

Table C: Order effects on trust

	Trust (%)		
	1st Game		**2nd Game**
BL-EXO-VOTE	41.7%	~	43.8%
	~		$\vee^{***}$
EXO-BL-VOTE	35.4%	$>^{**}$	14.6%

Notes: BL stands for *Baseline* treatment, EXO for *Exogenous* treatment, VOTE for *Voting* treatment. *** indicates the significance at 1% level based on logit regressions. The results are confirmed using two-tailed z-tests.

Supplementary Material -- Not for publication

Appendix 2: Experimental instructions (Baseline first + Information)

Welcome. This is a study on how people make decisions. In this study you can earn money based on how well you follow the instructions, and on the decisions made by you and by the other participants. You will be paid in private and in cash at the end of the session.
Please turn off your mobile phone. From this moment on, no form of communication among participants is allowed. If you have any question, or need assistance of any kind, please raise your hand and one of us will come to your desk to help you.
Please, follow the instructions carefully. In this study there are four parts, and for each part, we will distribute and read the corresponding instructions. In the first three parts, after having read the instructions, we will ask you to answer six questions, to verify your full understanding. For every question you answer correctly you earn €0.40. So you can earn up to €7.2 by answering correctly to all questions for Parts 1, 2 and 3 of the study. In addition you will earn money for the decisions you and the other participants will make in Parts 1, 2, 3 and 4 of the study.

Now, I will read instruction for Part 1.

Instructions for Part 1

In this part of the study, participants are randomly divided into **groups of six**. In each group, three participants will be assigned the role BLUE, while the other three will be RED, then the computer will form pairs of subjects belonging to the same group. If you are BLUE, you will be paired with a RED player, and vice versa. Your counterpart will never know your true identity, nor will you know hers/his.

Your earnings are expressed in tokens that will be converted in Euros at the rate of 1 Euro for 3 tokens.

BLUE has to make one choice: between option A and option B. RED has to make one choice: between option X and option Y. Table 1 summarizes the earnings corresponding to BLUE's and RED's choices.

Table 1: earnings in Part 1

BLUE chooses	RED chooses	Earnings
A	X	BLUE: 0 RED: 30
	Y	BLUE: 15 RED: 15
B	Irrelevant	BLUE: 8 RED: 8

If BLUE chooses option A, earnings depend on the choice made by RED:

- if RED chooses X, BLUE earns **0** tokens and RED earns **30** tokens;
- if RED chooses Y, BLUE earns **15** tokens and RED earns **15** tokens.

If BLUE chooses option B, the choice made by RED has no consequences on either BLUE's or RED's earnings:

- BLUE earns **8** tokens and RED earns **8** tokens.

We ask you to make a decision first as RED, then as BLUE. **We will inform you of the role you are actually assigned in this Part only at the end of the session.**

If you are assigned the BLUE role, your earnings from this part will depend on the choice you made as BLUE, and on the choice made by your counterpart as RED.

If you are assigned the RED role, your earnings from this part will depend on the choice you made as RED, and on the choice made by your counterpart as BLUE.

You will be informed of the results of this Part only at the end of the session.

We will now make an **example**. At the end of the example we will ask you to answer two questions, to verify you understanding of the instructions. Remember that you earn €0.40 for each question you answer correctly.

Look at your screen. You now have to make a choice as RED. Please, choose X, and confirm your choice. Good. You now have to make a choice as BLUE. Please, choose B and confirm your choice. Good. On your screen, you will now see two questions. Please, give your answers by pressing the corresponding buttons.

If you are not sure about the answer, you can re-read the instructions. Take your time and think carefully before answering the question.

[As RED, you chose X and as BLUE you chose B. You are assigned the BLUE role, and your counterpart, who is assigned the RED role, chose Y.

- *How much do you earn?*
- *How much does your counterpart earn?]*

We will now make another **example**. At the end of the example we will ask you to answer two questions, to verify you understanding of the instructions. Remember that you earn €0.40 for each question you answer correctly.

Look at your screen. You now have to make a choice as RED. Please, choose X, and confirm your choice. Good. You now have to make a choice as BLUE. Please, choose A and confirm your choice. Good. On your screen, you will now see two questions. Please, give your answers by pressing the corresponding buttons.

[As RED, you chose X and as BLUE you chose A. You are assigned the RED role, and your counterpart, who is assigned the BLUE role, chose A.

- *How much do you earn?*
- *How much does your counterpart earn?]*

You will now read on your screen the last two questions. Please, give your answers by pressing the corresponding buttons.

- *How much are 6 tokens worth, in Euros?*
- *Will you know if you are RED or BLUE before making your choice?*

If you have any doubts on the instructions, please raise your hand now. Good, then we can start with Part 1.

Instructions for Part 2

In this part of the study, participants are in **the same groups of six as in Part 1**. In each group, three participants will be assigned the role BLUE, while the other three will be RED, then the computer will form pairs of subjects belonging to the same group. If you are BLUE, you will be paired with a RED player, and vice versa. Your counterpart will never know your true identity, nor will you know hers/his. Your counterpart will **NOT** be the same person as in Part 1.
Your earnings are expressed in tokens that will be converted in Euros at the rate of 1 Euro for 3 tokens. You may also lose tokens. In the unlikely event your total earnings at the end of the study are negative, you may lose part of the money you earned by correctly answering the questions on the instructions. In any case, we guarantee you a minimum earning of €5 for your participation.
BLUE has to make one choice: between option A and option B. RED has to make one choice: between option X and option Y. Table 2 summarizes the earnings corresponding to BLUE's and RED's choices. *Earnings for RED may depend on the choices made by the other five members of the group.*

Table 2: earnings in Part 2

BLUE chooses	RED chooses	Earnings
A	X	BLUE: 0 RED: 30 – 8 x *number of others who choose Y*
	Y	BLUE: 15 RED: 15
B	Irrelevant	BLUE: 8 RED: 8

If BLUE chooses option A, earnings depend on the choice made by RED:

- if RED chooses X, BLUE earns **0** tokens. Earnings for RED depend on the choices made as RED by the other five members of the group. Notice that all members of your group make decisions both as RED and as BLUE, before knowing the role they are actually assigned.
    - If 0 of the others chooses Y, RED will get 30 tokens.
    - If 1 of others chooses Y, RED will get 22 tokens.
    - If 2 of others choose Y, RED will get 14 tokens.
    - If 3 of others choose Y, RED will get 6 tokens.
    - If 4 of others choose Y, RED will lose 2 tokens.
    - If 5 of others choose Y, RED will lose 10 tokens.
- if RED chooses Y, BLUE earns **15** tokens and RED earns **15** tokens.

If BLUE chooses option B, the choice made by RED has no consequences on either BLUE's or RED's earnings:

- BLUE earns **8** tokens and RED earns **8** tokens.

We ask you to make a decision first as RED, then as BLUE. **We will inform you of the role you are actually assigned in this Part only at the end of the session.**

If you are assigned the BLUE role, your earnings from this part will depend on the choice you made as BLUE, and on the choice made by your counterpart as RED.

If you are assigned the RED role, your earnings from this part will depend on the choice you made as RED, on the choice made by your counterpart as BLUE, *and on the choices made as RED by each of the other five members of your group*.

You will be informed of the results of this Part only at the end of the session.

We will now make an **example**. At the end of the example we will ask you to answer two questions, to verify you understanding of the instructions. Remember that you earn €0.40 for each question you answer correctly.

Look at your screen. You now have to make a choice as RED. Please, choose Y, and confirm your choice. Good. You now have to make a choice as BLUE. Please, choose B and confirm your choice. Good. On your screen, you will now see two questions. Please, give your answers by pressing the corresponding buttons.

If you are not sure about the answer, you can re-read the instructions. Take your time and think carefully before answering the question.

[As RED, you chose Y and as BLUE you chose B. You are assigned the BLUE role, and your counterpart, who is assigned the RED role, chose X. Two of the other members of your group chose Y as RED.

- *How much do you earn?*

- *How much does your counterpart earn?]*

We will now make another **example**. At the end of the example we will ask you to answer two questions, to verify you understanding of the instructions. Remember that you earn €0.40 for each question you answer correctly.

Look at your screen. You now have to make a choice as RED. Please, choose X, and confirm your choice. Good. You now have to make a choice as BLUE. Please, choose A and confirm your choice. Good. On your screen, you will now see two questions. Please, give your answers by pressing the corresponding buttons.

[As RED, you chose X and as BLUE you chose A. You are assigned the RED role, and your counterpart, who is assigned the BLUE role, chose A. Four of the other members of your group chose Y as RED.

- *How much do you earn?*
- *How much does your counterpart earn?]*

You will now read on your screen the last two questions. Please, give your answers by pressing the corresponding buttons.

- *Can your counterpart in Part 2 be the same person as in Part 1?*
- *How many people are there in each group?*

If you have any doubts on the instructions, please raise your hand now. Good, then we can start with Part 2.

Instructions for Part 3

In this part of the study, participants are in **the same groups of six as in Parts 1 and 2**. In each group, three participants will be assigned the role BLUE, while the other three will be RED, then the computer will form pairs of subjects belonging to the same group. If you are BLUE, you will be paired with a RED player, and vice versa. Your counterpart will never know your true identity, nor will you know hers/his. Your counterpart will **NOT** be the same person as in Part 1 or in Part 2.

In Part 3, you will be asked to take 3 decisions. First you will have vote in favor of either Situation 1, or Situation 2. Then you will have to make a choice as RED and as BLUE, as in Parts 1 and 2.

Situation 1 is the situation you faced in Part 1 of this study, represented in Table 3.

Table 3: Situation1

BLUE chooses	RED chooses	Earnings

A	**X**	BLUE: 0 RED: 30
	Y	BLUE: 15 RED: 15
B	**Irrelevant**	BLUE: 8 RED: 8

Situation 2 is the situation you faced in Part 2 of this study, represented in Table 4.

Table 4: Situation 2

BLUE chooses	**RED chooses**	**Earnings**
A	**X**	BLUE: 0 RED: 30 – 8 x *number of others who choose Y*
	Y	BLUE: 15 RED: 15
B	**Irrelevant**	BLUE: 8 RED: 8

When all participants have casted their vote, you will be informed of how many of your group's members voted for Situation 1, of how many of your group's members voted for Situation 2, and of the outcome of the vote.

If the majority of the members of your group vote for Situation 1, then the rules for the rest of this Part will be the same as in Part 1. If instead the majority of the members in your group vote for Situation 2, then the rules for the rest of this Part will be the same as in Part 2. If in your group three members vote in favor of Situation 1, and three members vote in favor of Situation 2, then the outcome will be randomly determined by the computer.

We ask you to make a decision first as RED, then as BLUE. **We will inform you of the role you are actually assigned only at the end of the session.**

If you are assigned the BLUE role, your earnings from this part will depend on the choice you made as BLUE, and on the choice made by your counterpart as RED.

If you are assigned the RED role, your earnings from this part will depend on the choice you made as

RED, and on the choice made by your counterpart as BLUE. *In case in your group the outcome of the vote is Situation 2, earnings for RED may also depend on the choices made as RED by each of the other five members of your group*.

You will be informed of the results of this Part only at the end of the session.

We will now make an **example**. At the end of the example we will ask you to answer two questions, to verify you understanding of the instructions. Remember that you earn €0.40 for each question you answer correctly.

Look at your screen. You now have to vote either for Situation 1 or for Situation 2. Please, vote for Situation 2, and confirm your choice.

You can now see on your screen that the majority of your group members voted for Situation 1. Hence, the rules for the rest of this Part will be the same as in Part 1.

You now have to make a choice as RED. Please, choose Y, and confirm your choice. Good. You now have to make a choice as BLUE. Please, choose B and confirm your choice. Good. On your screen, you will now see two questions. Please, give your answers by pressing the corresponding buttons.

If you are not sure about the answer, you can re-read the instructions. Take your time and think carefully before answering the question.

[Situation 1 has been selected. As RED, you chose Y and as BLUE you chose B. You are assigned the BLUE role, and your counterpart, who is assigned the RED role, chose X. Four of the other members of your group chose Y as RED.

- *How much do you earn?*
- *How much does your counterpart earn?]*

We will now make another **example**. At the end of the example we will ask you to answer two questions, to verify you understanding of the instructions. Remember that you earn €0.40 for each question you answer correctly.

Look at your screen. You now have to vote either for Situation 1 or for Situation 2. Please, vote for Situation 1, and confirm your choice.

You can now see on your screen that the majority of your group members voted for Situation 2. Hence, the rules for the rest of this Part will be the same as in Part 2.

You now have to make a choice as RED. Please, choose X, and confirm your choice. Good. You now have to make a choice as BLUE. Please, choose A and confirm your choice. Good. On your screen, you will now see two questions. Please, give your answers by pressing the corresponding buttons.

[Situation 2 has been selected. As RED, you chose X and as BLUE you chose A. You are assigned the

RED role, and your counterpart, who is assigned the BLUE role, chose A. Two of the other members of your group chose Y as RED.

- *How much do you earn?*
- *How much does your counterpart earn?]*

You will now read on your screen the last two questions. Please, give your answers by pressing the corresponding buttons.

- *Can your counterpart in Part 3 be the same person as in Part 1 or Part 2?*
- *If four members of your group vote for Situation 1 and two members of your group vote for Situation 2, in Part 3 your group will play according to the rules adopted in Part 1 of the study. True or False?*

If you have any doubts on the instructions, please raise your hand now. Good, then we can start with Part 3.

Instructions for Part 4

In this part of the study, participants are in **the same groups of six as in Parts 1, 2 and 3**. In each group, three participants will be assigned the role BLUE, while the other three will be RED, then the computer will form pairs of subjects belonging to the same group. If you are BLUE, you will be paired with a RED player, and vice versa. Your counterpart will never know your true identity, nor will you know hers/his. Your counterpart **may** be the same person as in **Part 1, Part 2 or in Part 3**.
Rules for Part 4 are the same as for Part 3: you will be asked to take 3 decisions. First you will have vote in favor of either Situation 1, or Situation 2. Then you will have to make a choice as RED and as BLUE, as in Parts 1, 2 and 3. **Differently from Part 3**, in Part 4, before making your decisions, you will receive **information on the choices** that you and your group members made in **Parts 1, and 2**.
At the end of this Part, you will receive information on the outcome of Parts 1, 2 3 and 4 of the study. You will know the role you have been assigned in each Part, and the earnings you obtained.

Appendix 3: Questionnaire

We kindly ask you to complete this questionnaire. The answers you give will not affect in any way your earnings. Some of these questions refer to personal information, which will help us in this study. Your identity will not be revealed under any circumstances in the presentation of the results.

Please answer carefully. Once an answer is given, you can no longer change it.

Press OK to begin. Thank you.

1. Were the instructions you have received for today's activities clear?

(1) No, not at all (2) No, not so much (3) Yes, enough (4) Yes, very much

2. Gender (press the corresponding button)

(1) Male (2) Female

3. Age (please, give your answer using the slider below and press ok to confirm)

4. Were you born in Italy?

(1) Yes (2) No

5. Education background

(1) Middle high school (2) High school (3) Bachelor degree
(4) Master degree (5) Ph.D. or postgraduate degree (6) Other

6. Occupation

(1) Student (2) Self-employed worker (3) Employee (4) Retired
(5) Jobless (6) Others

6.1 Field of studies (this question is accessed only if the subject gives answer (1) to question 6)

(1) Social sciences (2) Mathematical, Physical and Natural sciences
(3) Engineering and Architecture (4) Medicine
(5) Literature and Philosophy (6) Others

7. Have you attended courses in Economics?

(1) Yes (2) No

8. Have you attended courses in Statistics?

(1) Yes (2) No

9. Have you attended courses in Game Theory?

(1) Yes (2) No

10. Have you previously participated as a volunteer in other researches?
(choose one or more answers)

(1) Yes, in the field of economics
(2) Yes, in the field of psychology
(3) Yes, in the field of medicine or biology
(4) No

11. Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?

(1) Most people can be trusted(2) Can't be too careful (3) No idea

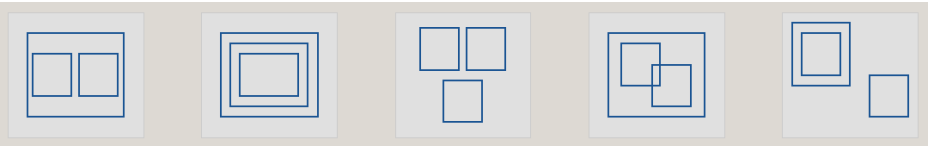
12. Are you generally a person who is fully prepared to take risks or do you try to avoid taking risk?
Please tick a box on the scale, where the value 1 means: "unwilling to take risks" and the value 10 means: "fully prepared to take risk"

1, 2, 3, 4, 5, 6, 7, 8, 9, 10

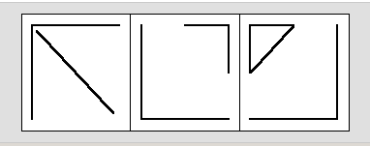
13. In general, do you think it is important to help others, and take care of their well being?
Please tick a box on the scale, where the value 1 means: "not important at all" and the value 10 means: "Maximally important"

1, 2, 3, 4, 5, 6, 7, 8, 9, 10

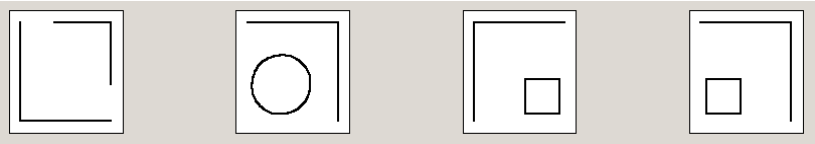
14. Which of these diagrams represents the relationship between Orange-Citrus Fruit-Fruit? Please select an answer and click OK to confirm.



15. Select the element that completes the following series.



Please select an answer and click OK to confirm.



16. A bat and a ball cost $ 1.10 in total. The bat costs $ 1.00 more than the ball. How much does the ball cost?

17. If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?

18. In a pond, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire pond, how long would it take for the patch to cover half of the pond?